KODITEX — KORPUS DIVERZIFIKOVANÝCH TEXTŮ

V listopadu 2018 vyšel z dílny Českého národního korpusu (ČNK; https://korpus.cz/) nový korpus češtiny — Koditex (Zasina, Lukeš, Komrsková, Poukarová & Řehořková, 2018). Jedná se o synchronní, reprezentativní a referenční korpus, který obsahuje 10,9 mil. pozic, což po odečtení interpunkce činí přibližně 9 mil. slov. Korpus byl vytvořen za účelem provedení multidimenzionální analýzy češtiny (Cvrček et al., 2018a, 2018b), proto se v mnoha ohledech od ostatních korpusů ČNK liší. Název korpusu je zkratkovým slovem pro *korpus diverzifikovaných textů* a současně odkazuje k osobě Viléma Kodýtka, který se jako první pokusil aplikovat multidimenzionální analýzu na češtinu (Kodýtek, nepubl.).

Při sestavování korpusu byl kladen důraz na co nejpestřejší složení, které by odráželo variabilitu současné[1] češtiny ve třech komunikačních módech: psaný, mluvený, internetový.[2] Snažili jsme se tedy vybrat co nejširší spektrum textů všech tří módů od co největšího počtu autorů/mluvčích. Právě s ohledem na zvýšení pestrosti vstupoval do korpusu každý text v podobě vzorku (*chunk*),[3] na rozdíl od ostatních synchronních korpusů ČNK (např. SYN2015), ve kterých jsou obsaženy celé texty. Celkově je Koditex tvořen 3 428 textovými vzorky, které zahrnují přes 719 tis. vět; obsahuje 506 764 slovních tvarů a 205 592 tisíc lemmat.

Koditex je strukturován hierarchicky (viz tab. 1). Nejvyšší úroveň tvoří komunikační módy (*mode*): psaný jazyk (*wri*), mluvený jazyk (*spo*) a internetová komunikace (*web*). Každý mód se dále dělí do dvou a více divizí (*division*). Divize jsou rozděleny do tříd (*class*), jejichž cílová velikost činí přibližně 200 000 slov na třídu. Psaný mód vyžadoval zavedení mezistupně, tzv. nadtřídy (*superclass*), seskupující některé třídy. Jak struktura, tak hierarchická klasifikace textů do tříd vychází z dřívějších zkušeností s výstavbou korpusů ČNK (srov. Cvrček, Čermáková & Křen, 2016).

Zdroje dat pro jednotlivé textové třídy se liší. Většina textů pochází z vlastních zdrojů ČNK. Texty, jimiž ČNK nedisponoval, byly poskytnuty jinými vědecko-výzkumnými pracovišti (viz dále). Textové vzorky v psaném módu pocházejí převážně z korpusu SYN2015 (Křen et al., 2015), pouze v některých případech bylo k naplnění třídy nutno využít i jiné korpusy řady SYN. Pro třídu soukromé korespondence (*class: pri*) jsme využili Korpus soukromé korespondence (Hladká, 2006). V mluveném módu jsou použity pro každou třídu jiné zdroje: neformální komunikace (*class: inf*) pochází z korpusu ORAL2013 (Benešová, Křen & Waclawičová, 2013), elicitovaná komunikace

1 Korpus obsahuje texty, jež vznikly po roce 1990; většina z nich pochází z let 2007–2015.

2 Vydělení internetové komunikace do samostatného módu (podobně i Herringová, 2001) bylo motivováno především dosud chybějícím uspokojivým popisem textové složky českého internetu (srov. Biber & Egbert, 2016) a větší selekcí těchto dat při vytváření korpusu (při výběru byly uvažovány např. pouze psané texty, ačkoli lze uvažovat o řadě mluvených žánrů specifických pouze pro internet, pouze texty s diakritikou apod.). Internetové texty se od textů ve zbylých dvou módech navíc odlišují metadaty, mezi kterými dominují (spíše) technické údaje vztahující se k textu (např. denní doba, ve které byl text zveřejněn) oproti jedinému údaji o autorovi — jeho jménu, resp. přezdívce.

3 V závorkách uvádíme výraz v podobě, v jaké se objevuje při práci s korpusem.

MÓD (mode)	DIVIZE (division)	NADTŘÍDA (superclass)	TŘÍDA (class)	Slova	Textové vzorky
spo (mluvená komunikace)	int (interaktivní)		bru (nepřipravené veřejné/vysílané rozhovory)	221 812	90
			eli (formální rozhovor)	201 690	82
			inf (neformální rozhovor)	208 565	86
	nin (neinteraktivní)		wbs (připravený/čtený projev)	213 201	71
web (internetová komunikace)	mul (mnohosměrná komunikace)		dis (internetové diskuse)	197 948	87
			fcb (facebookové statusy)	199 418	91
			for (webová fóra)	200 104	85
	uni (jednosměrná komunikace)		blo (blogy)	204 356	74
			wik (wikipedie)	201 691	84
wri (psaná komunikace)	fic (beletrie)	nov (romány)	crm (detektivky)	190 026	68
			fan (fantasy)	189 432	69
			gen (bez bližšího určení)	193 667	67
			lov (milostné)	189 893	70
			scf (sci-fi)	188 703	68
			col (povídky)	195 595	70
			scr (scénáře a dramata)	182 689	76
			ver (poezie a písně)	205 837	76
	nfc (oborová literatura)	pop (populárně naučná)	fts (formální a technické vědy)	207 607	68
			hum (humanitní vědy)	204 837	74
			nat (přírodní vědy)	204 751	71
			ssc (společenské vědy)	203 698	68
		pro (profesní literatura)	fts (formální a technické vědy)	210 010	71
			hum (humanitní vědy)	207 916	69
			nat (přírodní vědy)	209 580	70
			ssc (společenské vědy)	209 385	72
		sci (vědecká literatura)	fts (formální a technické vědy)	202 932	67
			hum (humanitní vědy)	204 300	71
			nat (přírodní vědy)	206 716	72
			ssc (společenské vědy)	205 358	67
			adm (administrativa)	207 748	90
			enc (encyklopedie)	203 957	73
			mem (auto-/biografie)	203 390	71

MÓD (mode)	DIVIZE (division)	NADTŘÍDA (superclass)	TŘÍDA (class)	Slova	Textové vzorky
wri (psaná komunikace)	nmg (noviny a časopisy)	lei (volnočasová publicistika)	hou (bydlení, zahrada, hobby)	207 499	68
			int (zajímavosti ze světa)	209 232	69
			lif (životní styl)	203 124	72
			mix (víkendové přílohy)	205 310	75
			sct (bulvár)	201 417	73
			spo (sport)	199 238	70
		new (tradiční publicistika)	com (komentáře)	205 372	68
			cul (kultura)	205 690	68
			eco (ekonomika)	211 481	70
			fre (volnočasové aktivity)	208 532	71
			pol (politika)	206 893	70
			rep (reportáže)	206 377	70
	pri (soukromá komunikace)		cor (dopisy)	195 321	196
Celkem				9 142 298	3428

TABULKA 1: Složení korpusu Koditex.[4]

(*class: eli*) z formální části korpusů PMK (Čermák, Adamovičová & Pešička, 2001) a BMK (Hladká, 2002), mediální komunikace (*class: bru*) z korpusu DIALOG (http://ujc.dialogy.cz/),[5] připravené a převážně čtené projevy (*class: wbs*) z korpusu SPEECHES (Cvrček, Truneček & Horký, 2015) a z české části zápisů v korpusu EUROPARL (http://www.statmt.org/europarl/). Texty v módu internetové komunikace jsme také vybírali z několika zdrojů: vzorky z internetové encyklopedie Wikipedie (*class: wik*) z Korpusu české Wikipedie, sestaveného Centrem zpracování přirozeného jazyka v Brně,[6] blogy (*class: blo*) z korpusu Araneum Bohemicum Maius (Benko, 2015) a data pro kategorie příspěvků z diskusních fór (*class: for*), internetových diskusí (*class: dis*) a veřejných statusů ze sociální sítě Facebook (*class: fac*) poskytl Josef Šlerka a společnost Socialinsider.

Většina textů (s pokrytím 76 % všech tokenů) zahrnutých v korpusu představuje české originály (tedy nikoliv překlady z jiných jazyků). Jedinou výjimkou jsou textové třídy, u kterých je v češtině výskyt přeložených materiálů zcela běžný. Ve třídách *lov*, *crm*, *gen*, *fan*, *scf*, *mem* tvoří podíl překladů více než 70 %, ve třídách *hum*, *nat*, *enc*, *ssc*,

4 Texty původně zařazené do korpusu bylo v některých případech třeba z analýzy registrové variability vyloučit. V takovém případě mají v metadatech příznak include="no". Tato tabulka ukazuje složení celého korpusu Koditex.

5 Za poskytnutá data děkujeme Martinu Proškovi a Petru Kaderkovi z Ústavu pro jazyk český AV ČR.

6 Za poskytnutí dat děkujeme Karlu Palovi a Vítu Baisovi z Centra zpracování přirozeného jazyka na Masarykově univerzitě.

fts, *ver*, *wik* tvoří překlady 20–45 %. V ostatních třídách jsou zahrnuty české originály ze 100 %.

V procesu vytváření korpusu Koditex jsme v první fázi — vzorkování textů — stanovili délku textového vzorku na 2000 až 5000 slov se zachováním hranic vět. Texty delší než 5000 slov tedy bylo nutno rozdělit na menší části (z nichž do korpusu vstoupila většinou pouze jedna). Naopak některé textové třídy ukázaly, že obsahují texty zpravidla kratší než 2000 slov. Takové případy jsme řešili následovně: ve třídách administrativa a soukromá korespondence jsme snížili limit na 1000 slov a ve třídách obsahujících krátké příspěvky z webu, tj. třídy *for*, *dis*, *fac*, jsme spojili texty od téhož autora do jednoho vzorku o velikosti 2000–5000 slov. U transkriptů mluveného jazyka jsme analogicky spojovali repliky od téhož mluvčího hovořícího v rámci jedné nahrávky do jednoho celku, takže souhrny replik od jednotlivých komunikačních partnerů vstupovaly do korpusu jako samostatné komunikáty.[7]

Pro některé třídy jsme měli k dispozici více vzorků, než bylo potřeba, proto jsme stanovenou kvótu 200 000 slov naplňovali na základě algoritmu vytvářejícího co nejpestřejší směs textů.[8] Algoritmus v rámci jedné třídy nejdříve vybral náhodně první vzorek a následně se snažil najít další vzorek, který se od předchozího lišil v co největším počtu relevantních metainformací.[9] Takový vzorek byl přidán k výběru a algoritmus pokračoval v hledání až do naplnění požadované velikosti.

Korpus Koditex je anotován na několika úrovních. Kromě standardní anotace (lemmatizace, morfologické značkování[10]) obsahuje anotaci frazémů pomocí systému FRANTA (Hnátková, 2002) a tzv. pojmenovaných entit (*named entities*) nástrojem NameTag (http://ufal.mff.cuni.cz/nametag; Straková, Straka & Hajič, 2013).

Korpus Koditex představuje specializovaný typ jazykového korpusu, který je v rámci korpusů ČNK unikátní. Jeho specifičnost tkví především v účelu, za jakým byl vytvořen, od něhož se odvíjela i jeho neobvyklá struktura a mnohovrstevnatá anotace. Provedení multidimenzionální analýzy registrové variability češtiny, jíž sloužil jako datová základna, však neznamená vyčerpání všech možností jeho využití. Domníváme se, že nové možnosti naopak otevírá. Jednotné zpracování dat ze psané, mluvené a internetové komunikace včetně zpřístupnění nových žánrů, např. příspěvků z internetových diskusí, umožňuje jejich srovnání na různých rovinách jazykového popisu. Doufáme, že korpus Koditex bude nejen inspirací ke vzniku nových studií.

7 Podrobný popis je uveden v Cvrček et. al. (2018a, s. 3–7).

8 Za koncepci a implementaci tohoto algoritmu vděčíme Jiřímu Václavíkovi.

9 V jednotlivých textových třídách se inventář dostupných metainformací lišil (např. informace o vzdělání autora byla k dispozici pouze u části mluvených dat). Obecně však algoritmus přihlížel ke všem charakteristikám textu či jeho autora, u nichž lze předpokládat vliv na jazykovou variabilitu (vyloučen tak byl např. údaj ISBN, který žádnou relevantní charakteristiku textu neodráží).

10 Morfologické značkování bylo provedeno dvěma systémy: stochastickým taggerem MorphoDiTa (http://ufal.mff.cuni.cz/morphodita; Straková, Straka & Hajič, 2014) a hybridním taggerem s použitím stochastické a pravidlové desambiguace (Spoustová, Hajič, Votrubec, Krbec & Květoň, 2007; Jelínek, 2008; Petkevič, 2014).

LITERATURA

Benešová, L., Křen, M., & Waclawičová, M. (2013). *ORAL2013: reprezentativní korpus neformální mluvené češtiny*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Benko, V. (2015). *Araneum Bohemicum Maius, verze 15.04*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Biber, D., & Egbert, J. (2016). Register Variation on the Searchable Web: A Multi-Dimensional Analysis. *Journal of English Linguistics*, *44*(2), 95–137.

Cvrček, V., Truneček, P., & Horký, V. (2015). *Korpus prezidentských projevů Speeches*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Cvrček, V., Čermáková, A., & Křen, M. (2016). Nová koncepce synchronních korpusů psané češtiny. *Slovo a slovesnost*, *77*(2), 83–101.

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018a): From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory*. Dostupné z https://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2018-0020/cllt-2018-0020.xml.

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018b). Variabilita češtiny: multidimenzionální analýza. *Slovo a slovesnost*, *79*(4), 293–321.

Čermák, F., Adamovičová, A. & Pešička, J. (2001). *PMK (Pražský mluvený korpus): přepisy nahrávek pražské mluvy z 90. let 20. století*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Herring, S. C. (2001). Computer Mediated Discourse. In D. Schiffrin, D. Tannen & H. Hamilton (Eds.), *The Handbook of Discourse Analysis* (s. 612–634). London, UK: Blackwell.

Hladká, Z. (2002). *BMK (Brněnský mluvený korpus): přepisy nahrávek brněnské mluvy z 90. let 20. století*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Hladká, Z. (2006). *KSK-dopisy (Korpus soukromé korespondence): přepisy ručně psaných dopisů z let 1990–2004*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Hnátková, M. (2002). Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. *Slovo a slovesnost*, *63*(2). 117–126.

Jelínek, T. (2008). Nové značkování v Českém národním korpusu. *Naše řeč*, *91*(1). 13–20.

Kodýtek, V. (nepublikováno). A translation of Biber's three-dimensional model of English into Czech. Dostupné z https://www.korpus.cz/biblio/2722.

Korpus DIALOG 1.1 (2012). Praha: Ústav pro jazyk český, AV ČR. Dostupné z http://ujc.dialogy.cz.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P. & Zasina, A. (2015). *SYN2015: reprezentativní korpus psané češtiny*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Petkevič, V. (2014). Problémy automatické morfologické disambiguace češtiny. *Naše řeč*, *97*(4), 194–207.

Spoustová, D., Hajič, J., Votrubec, J., Krbec, P. & Květoň, P. (2007). The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In J. Piskorski, B. Pouliquen, R. Steinberger & H. Tanev (Eds.), *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies* (s. 67–74). Praha: Association for Computational Linguistics.

Straková, J., Straka, M. & Hajič, J. (2013). A New State-of-The-Art Czech Named Entity Recognizer. In I. Habernal & V. Matoušek (Eds.), *Text, Speech and Dialogue* (s. 68–75). Berlin, Heidelberg: Springer Verlag.

Straková, J., Straka, M. & Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition.

In K. Bontcheva & Z. Jingbo (Eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (s. 13–18). Baltimore, MD: Association for Computational Linguistics.

Zasina, A. J., Lukeš, D., Komrsková, Z., Poukarová, P. & Řehořková, A. (2018). *Koditex: korpus diverzifikovaných textů*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Adrian Jan Zasina
Zuzana Komrsková